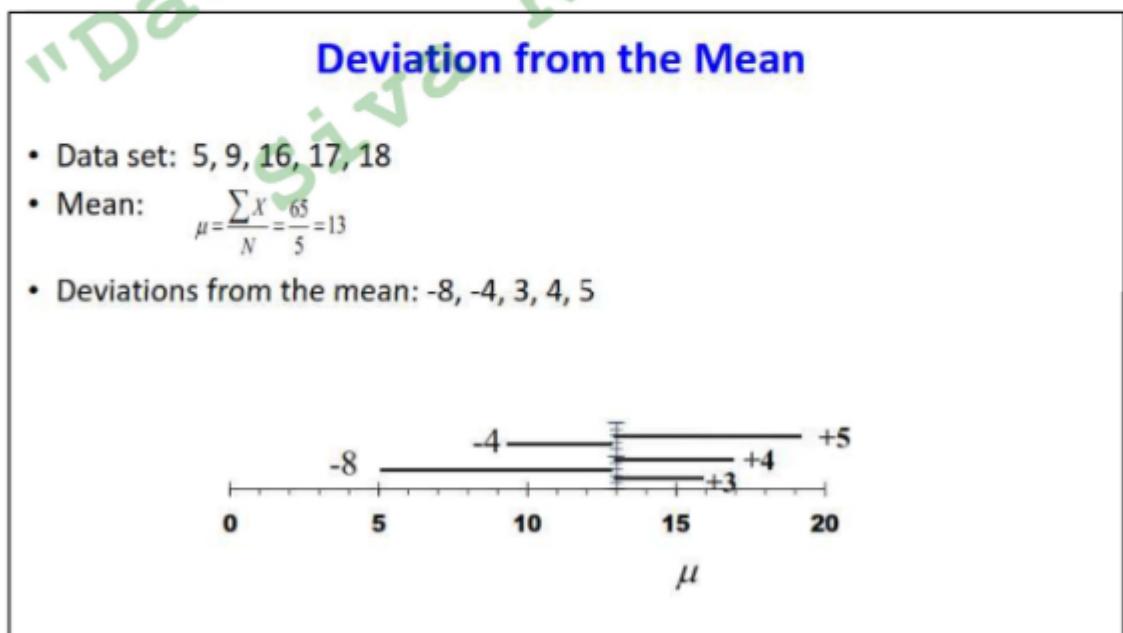
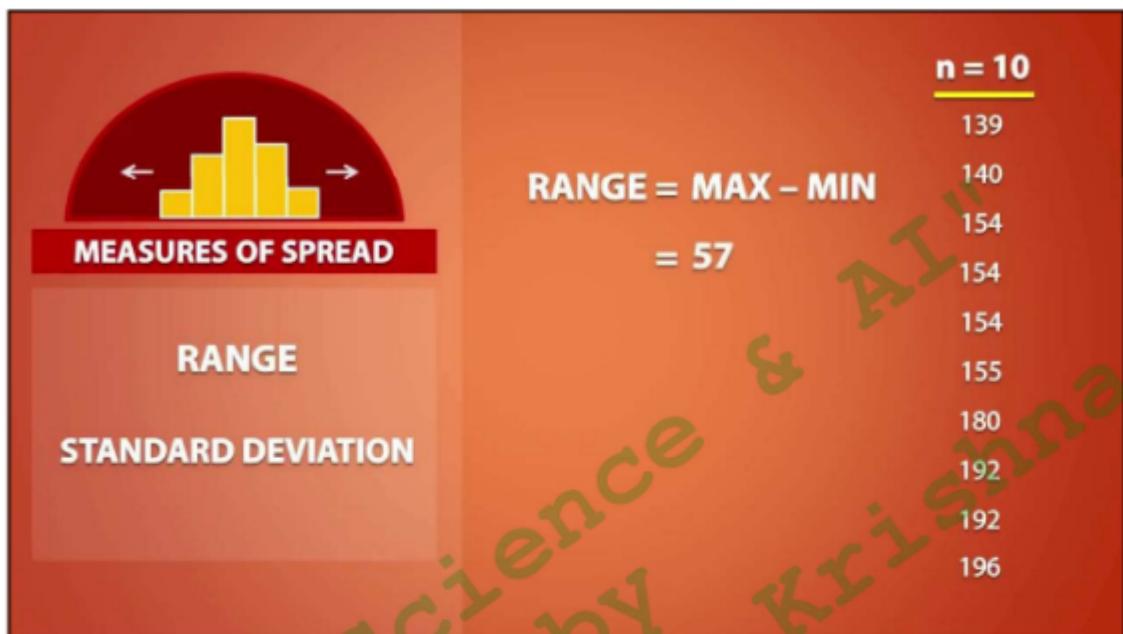




# MEASURES of DISPERSION / SPREAD







## Mean Absolute Deviation

- Average of the absolute deviations from the mean

$X$	$X - \mu$	$ X - \mu $
5	-8	+8
9	-4	+4
16	+3	+3
17	+4	+4
18	+5	+5
	0	24

$$M.A.D. = \frac{\sum |X - \mu|}{N}$$
$$= \frac{24}{5}$$
$$= 4.8$$

## Population Variance

- Average of the squared deviations from the arithmetic mean

$X$	$X - \mu$	$(X - \mu)^2$
5	-8	64
9	-4	16
16	+3	9
17	+4	16
18	+5	25
	0	130

$$\sigma^2 = \frac{\sum (X - \mu)^2}{N}$$
$$= \frac{130}{5}$$
$$= 26.0$$



## Population Standard Deviation

- Square root of the variance

X	X - μ	(X - μ) <sup>2</sup>
5	-8	64
9	-4	16
16	+3	9
17	+4	16
18	+5	25
	0	130

$$\begin{aligned}\sigma^2 &= \frac{\sum(X-\mu)^2}{N} \\ &= \frac{130}{5} \\ &= 26.0 \\ \sigma &= \sqrt{\sigma^2} \\ &= \sqrt{26.0} \\ &= 5.1\end{aligned}$$

### SAMPLE VARIANCE

$$s^2 = \frac{\sum (x_i - \bar{x})^2}{n-1}$$

### SAMPLE STANDARD DEVIATION

$$s = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}}$$



## Sample Variance

- Average of the squared deviations from the arithmetic mean

X	X - $\bar{X}$	$(X - \bar{X})^2$
2,398	625	390,625
1,844	71	5,041
1,539	-234	54,756
1,311	-462	213,444
7,092	0	663,866

$$S^2 = \frac{\sum(X - \bar{X})^2}{n-1}$$
$$= \frac{663,866}{3}$$
$$= 221,288.67$$

## Sample Standard Deviation

- Square root of the sample variance

X	X - $\bar{X}$	$(X - \bar{X})^2$
2,398	625	390,625
1,844	71	5,041
1,539	-234	54,756
1,311	-462	213,444
7,092	0	663,866

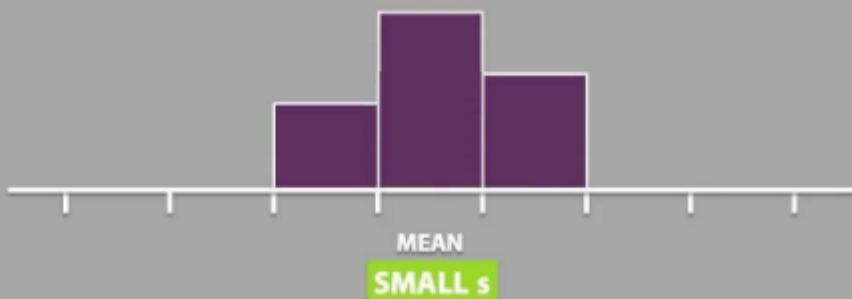
$$S^2 = \frac{\sum(X - \bar{X})^2}{n-1}$$
$$= \frac{663,866}{3}$$
$$= 221,288.67$$
$$S = \sqrt{S^2}$$
$$= \sqrt{221,288.67}$$
$$= 470.41$$



## STANDARD DEVIATION

HOW CLOSE THE VALUES IN A DATA SET  
ARE TO THE MEAN

STANDARD DEVIATION  
HOW CLOSE THE VALUES IN A DATA SET  
ARE TO THE MEAN





## STANDARD DEVIATION

HOW CLOSE THE VALUES IN A DATA SET  
ARE TO THE MEAN



## Coefficient of Variation

- Ratio of the standard deviation to the mean, expressed as a percentage
- Measurement of relative dispersion

$$C.V. = \frac{\sigma}{\mu} (100)$$



## Coefficient of Variation

$$\mu_1 = 29$$

$$\sigma_1 = 4.6$$

$$C.V_{.1} = \frac{\sigma_1}{\mu_1} (100)$$
$$= \frac{4.6}{29} (100)$$
$$= 15.86$$

$$\mu_2 = 84$$

$$\sigma_2 = 10$$

$$C.V_{.2} = \frac{\sigma_2}{\mu_2} (100)$$
$$= \frac{10}{84} (100)$$
$$= 11.90$$

"Data Science by Siva Rama Krishna"

**MEAS**

**C**

**CENTRAL**

*data science*

*data science*

*R*



154cm



## MEASURES OF CENTRE

**MODE = 154**

REFERS TO THE DATA VALUE  
THAT IS MOST FREQUENTLY  
OBSERVED



## MEASURES OF CENTRE

REFERS TO THE DATA VALUE  
THAT IS POSITIONED IN THE

**MEDIAN**

MIDDLE OF AN ORDERED  
DATA SET



## MEASURES OF CENTRE

REFERS TO THE DATA VALUE  
THAT IS POSITIONED IN THE

**MEDIAN**

MIDDLE OF AN ORDERED  
DATA SET

## Population Mean

$$\begin{aligned}\mu &= \frac{\sum X}{N} = \frac{X_1 + X_2 + X_3 + X_4 + X_5}{5} \\ &= \frac{24 + 13 + 19 + 26 + 11}{5} \\ &= \frac{93}{5} \\ &= 18.6\end{aligned}$$

## Weighted Average

- Sometimes we wish to average numbers according to their relative importance, or weight, to some other factor.
- The average you need is the **weighted average**.

*Weighted Average*

where  $x$  is a data value,  $w$  is the weight assigned to  $x$ , and  $n$  is the number of data values. The sum is the sum of all data values.

## Various Sampling Techniques

- **Random sampling**

- Every unit of the population has the same chance of being included in the sample.
- A chance mechanism is used in the selection process.
- Eliminates bias in the selection process.
- Also known as probability sampling

- **Nonrandom Sampling**

- Every unit of the population does not have an equal chance of being included in the sample.
- Open to the selection bias
- Not appropriate data collection method
- Also known as non-probability sampling

## Simple Random Sampling

- Every object in the population has an equal chance of being selected.
- Objects are selected independently of each other.
- Samples can be obtained from a table of random numbers or by using random number generators.
- A simple random sample is the ideal type of sample because all objects in the population are equally likely to be included and all samples of the same size have an equal chance of being selected.

## Systematic Sampling

- Convenient and relatively easy to administer
- Population elements are an ordered sequence (at least, conceptually).
- The first sample element is selected randomly from the first  $k$  population elements.
- Thereafter, sample elements are selected at a constant interval,  $k$ , from the ordered sequence frame.

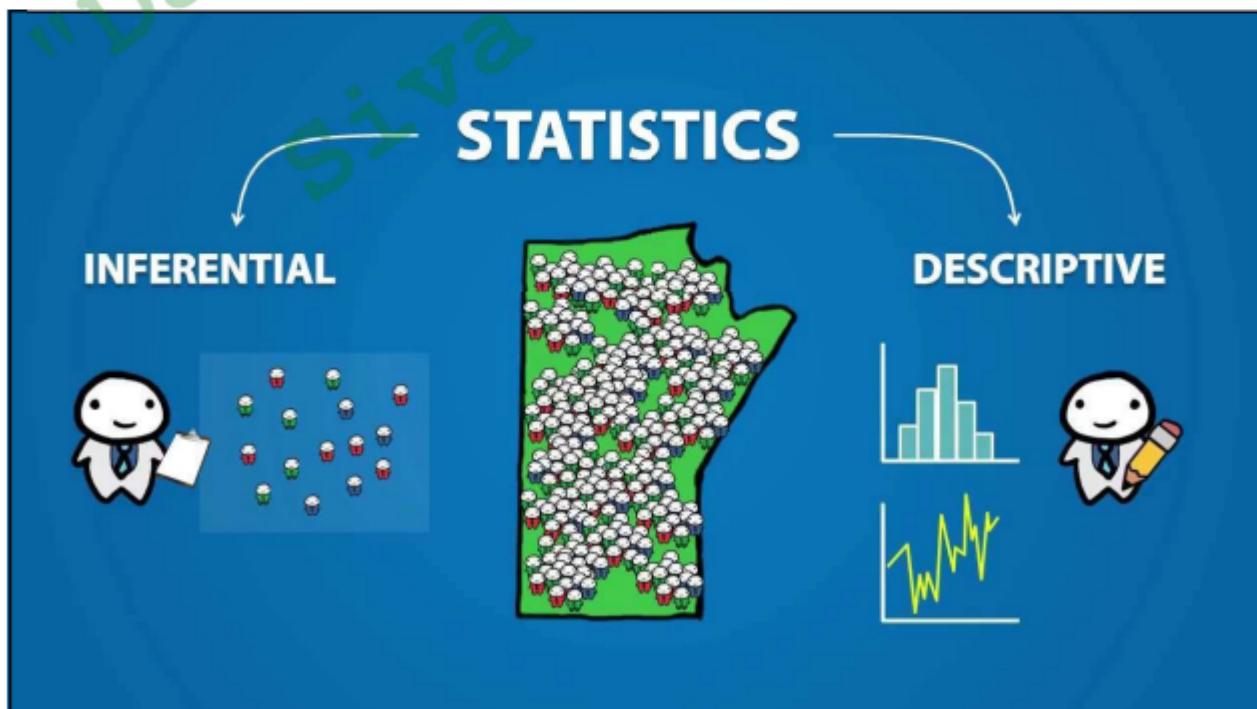
## Nonrandom Sampling

- **Convenience Sampling:** Sample elements are chosen based on convenience of the researcher
- **Judgment Sampling:** Sample elements are chosen based on the judgment of the researcher
- **Quota Sampling:** Sample elements are chosen until specific quotas are satisfied
- **Snowball Sampling:** Survey subjects are recruited from other survey respondents



# STATISTICS

MEASURE + ANALYZE





## Descriptive vs Inferential Statistics

- **Descriptive statistics**
  - Collecting, presenting, and describing data
- **Inferential statistics**
  - Drawing conclusions and/or making decisions concerning a population based only on sample data



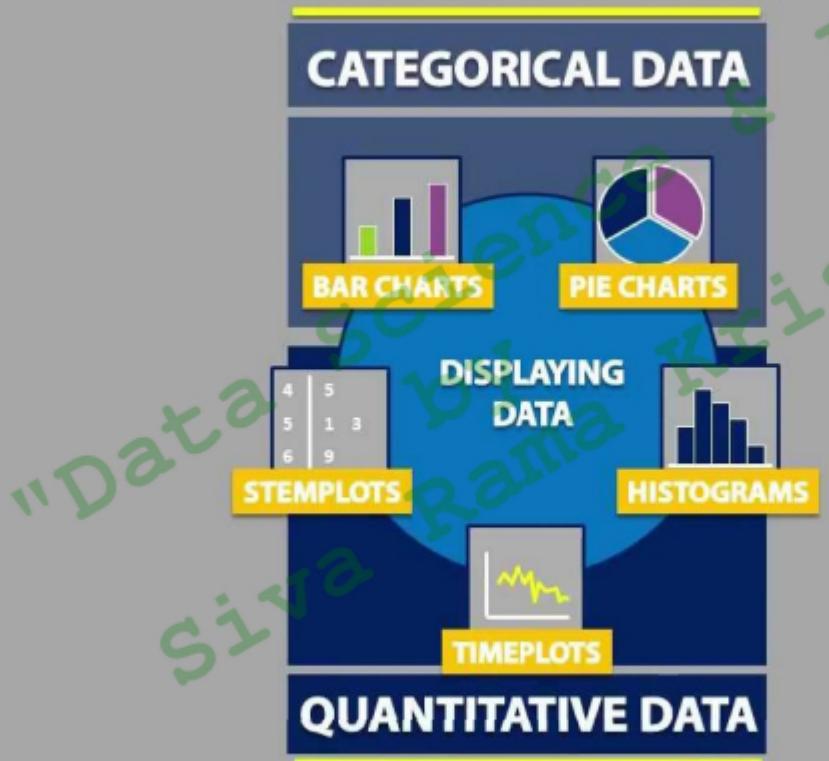
## Populations and Samples

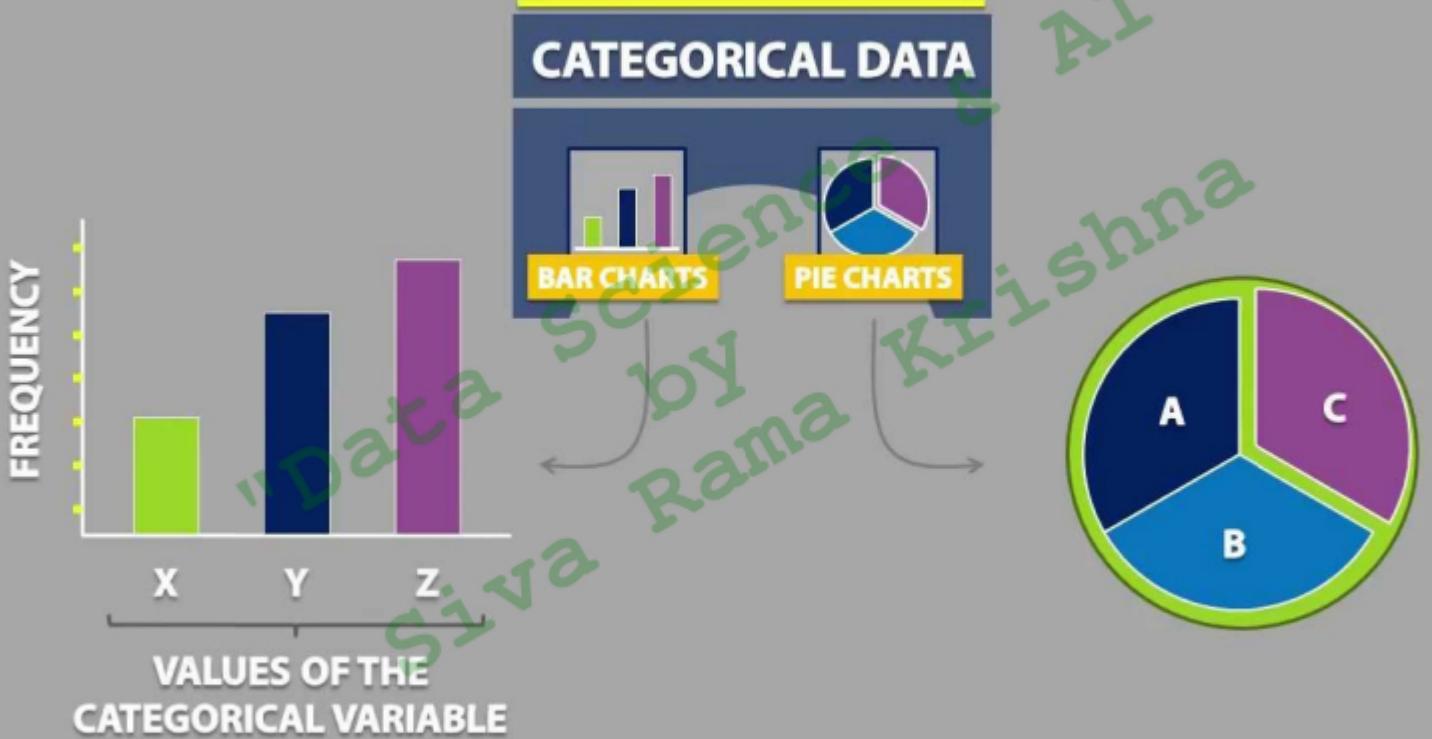
- A **Population** is the set of all items or individuals of interest
  - Examples: All likely voters in the next election  
All parts produced today  
All sales receipts for November
- A **Sample** is a subset of the population
  - Examples: 1000 voters selected at random for interview  
A few parts selected for destructive testing  
Random receipts selected for audit

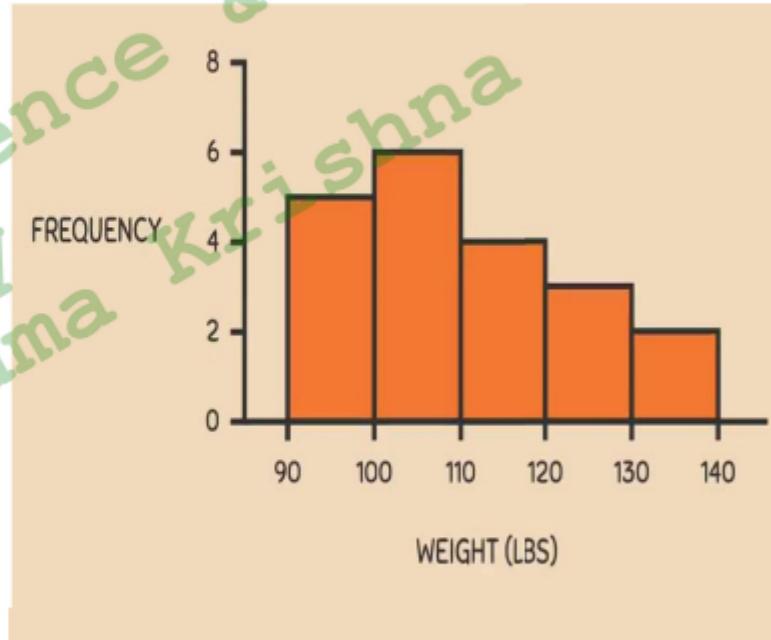
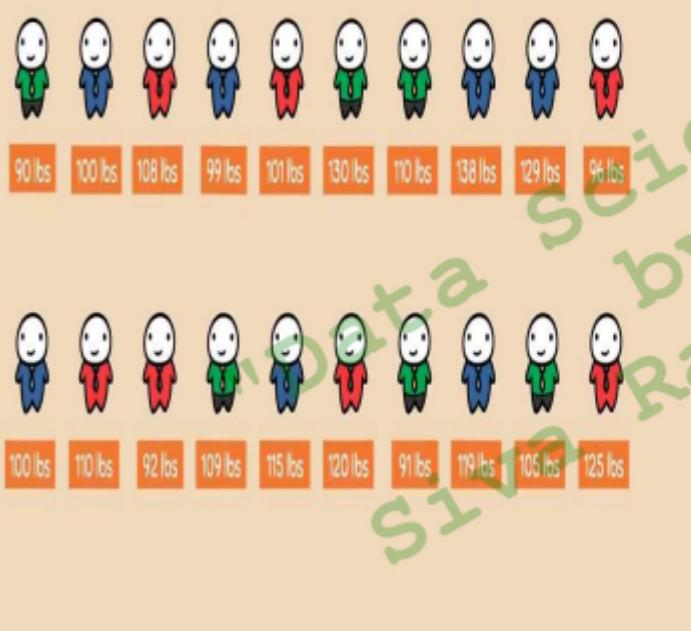


## Why Sample?

- Less time consuming than a census
- Less costly to administer than a census
- It is possible to obtain statistical results of a sufficiently high precision based on samples.
- Because the research process is sometimes destructive, the sample can save product
- If accessing the population is impossible; sampling is the only option









# FREQUENCY DISTRIBUTION



## WEIGHT

100 – 110	
110 – 120	120
120 – 130	130
130 – 140	
140 – 150	

BY CONVENTION, WE SAY THAT  
EACH INTERVAL DOES NOT  
INCLUDE THE RIGHT END POINT

- 
- 
-



# "Data Science & AI"

## Siva Ranjan Krishna

# TIMEPLOT

SHOW HOW A VARIABLE CHANGES OVER TIME



# TIMEPLOT

SHOW HOW A VARIABLE CHANGES OVER TIME

